Overview of State-Space Surveillance Modeling for Risk Analysis

Tom Ingersoll, Ph.D. [1], Christine D. Hoppe [2]

(1) Defense Threat Reduction Agency, (2) U.S. Army Edgewood Chemical Biological Center

Abstract

State-space modeling, sometimes known as hierarchical or mark-recapture modeling, is an innovative surveillance modeling tool that can correct for underreporting in surveillance data, estimate probability of false positives, and allow data fusion between partially-correlated data sources. Most threat agents cannot be perfectly detected under all conditions. Imperfect detection can cause underreporting bias, where observed threat conditions appear to be less severe than they actually are. Agents may go undetected although present on the landscape (false negatives), or may be detected at lower densities than are actually present (underestimates). Agents may be detected unevenly (observation heterogeneity) by different observers, different devices, or at different locations. Additionally, harmless agents may be mistaken for threat agents due to detection limitations (false positives). State-space modeling can correct for these sources of bias, improving the reliability of surveillance, sensing, risk analysis, and threat reporting.

There are two processes contributing to collected data, an observation process affected by interaction between observer and environment, and an underlying, partially-observed state process which is the density or distribution of the agent we wish to characterize. State-space models provide explicit estimates of both these processes. State-space models incorporate supplemental information from data that have multiple records at each spatial location, by either sampling repeatedly through time (temporally repeated measures: TRM), or by using multiple independent observers (MIO). It is the differences between repeated measures that allow us to separate estimates of detection and agent state to correct for underreporting. Other Chemical, Biological, Radiological and Nuclear (CBRN) surveillance modeling methods can underestimate risk when detection is less than perfect, or overestimate risk when false-positives occur.

State-space modeling uses existing data collection methods such as sensor networks, or disease surveillance records. The novelty of the method occurs during analysis, where observation and state are dissembled into separate model stages, then recombined to produce an integrated model likelihood. Analysis can occur within a frequentist or Bayesian framework, using software in the mathematical computing environment R or WINBUGS respectively.

State-space modeling has tremendous potential for enhancing surveillance in CBRN defense. Particularly, state-space models explicitly estimate a detection probability, which allows data-fusion, and the objective assessment of false-negatives and positives to be included in risk analysis. An overview of state-space modeling theory and methods is presented to an audience of intermediate mathematical literacy, that is, non-mathematicians with scientific or engineering training. State-space modeling literature is reviewed where data are relevant to CBRN surveillance, such as disease surveillance, and sensor data. State-space modeling methods are applied, as demonstration, to simulated data. TRM, and MIO are compared to estimates from non-repeated measures for precision and bias.

Formulation of a two-tiered state space model for observation and state

A model for distribution or incidence with under-reporting:

$$y_i \sim Bin(p_i z_i) \quad \} \text{ The observation model} \qquad \textit{Equation1}$$

$$z_i \sim Bern(\psi_i) \quad \} \text{ The state model} \qquad \textit{Equation 2}$$

Where y_i is the observed occurrence/non-occurrence at location *i*, *p* is the detection probability, z_i is the indicator function, ψ_i is the probability of incidence at site *i.*

Observation and/or state can vary from location to location through association with covariates:

$$\operatorname{logit}(\phi_i) = \sum_{0:n} \beta_n x_n \quad \} \text{ Heterogeneous state} \qquad \textit{Equation 3}$$

$$\operatorname{logit}(p_i) = \sum_{0:m} \alpha_m w_m \quad \} \text{ Heterogeneous detection} \qquad \textit{Equation 4}$$

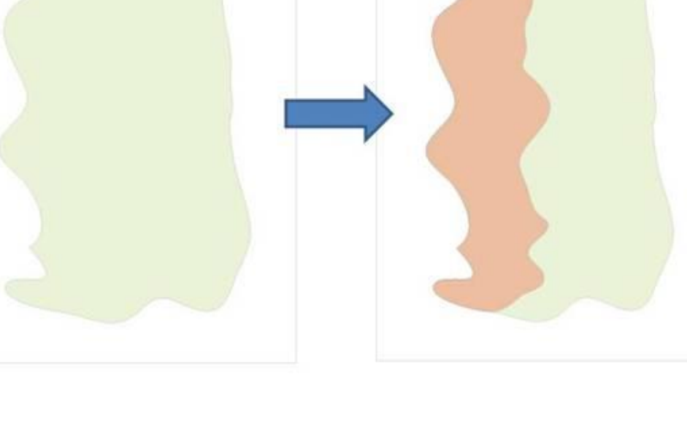
Where $\beta_{0:n}$ are incidence coefficients, $x_{1:n}$ are incidence covariates, $\alpha_{1:m}$ are detection coefficients, and $w_{1:m}$ are detection covariates.

Models for states other than incidence, such as abundance or density, may be formulated by substituting count or proportion distributions for the Bernoulli distribution in *Equation 2*.

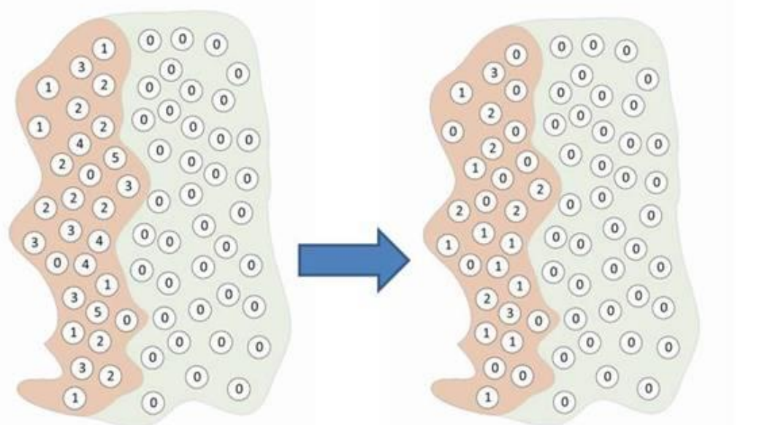
We use state-space modeling to correct for under-reporting in disease, and other surveillance data:

Underreporting and Zero-inflation (simulation)

Consider a region in the early stages of invasion by an exotic disease. The region will be divided into affected and non-affected zones:



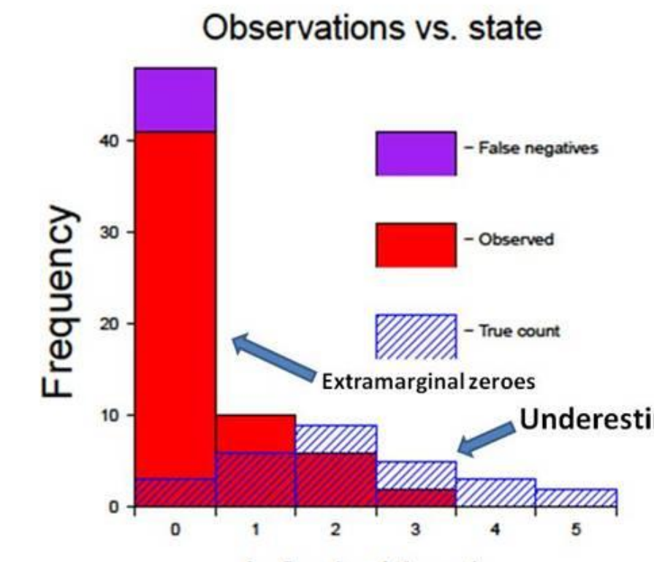
When infected individuals are imperfectly detected, field counts of the infected underestimate the true count:



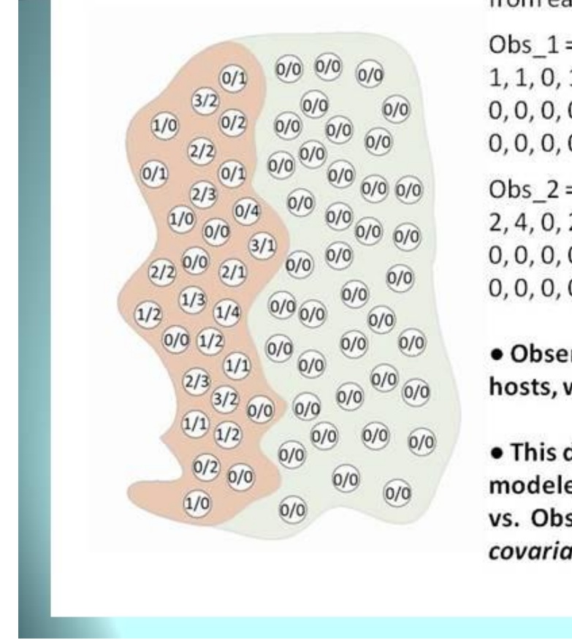
True population Under-reported count

Some affected locations will appear to be disease-free! (i.e. false negatives)

Observed surveillance data are a mixture of extra zeroes, and underestimated counts:



This can be remedied by sending a second observer to repeat the surveys:



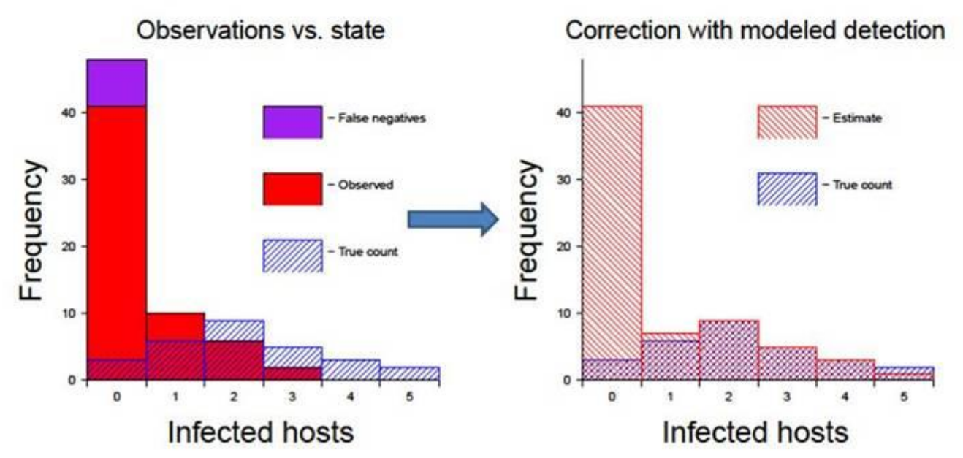
This results in 2 vectors of survey results, one from each observer:

Obs_1 = (0, 3, 1, 0, 2, 0, 0, 2, 1, 0, 0, 2, 2, 2, 1, 1, 1, 0, 1, 1, 2, 3, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

Obs_2 = (1, 2, 0, 2, 2, 1, 1, 3, 0, 4, 0, 1, 2, 2, 3, 2, 4, 0, 2, 1, 3, 2, 0, 1, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

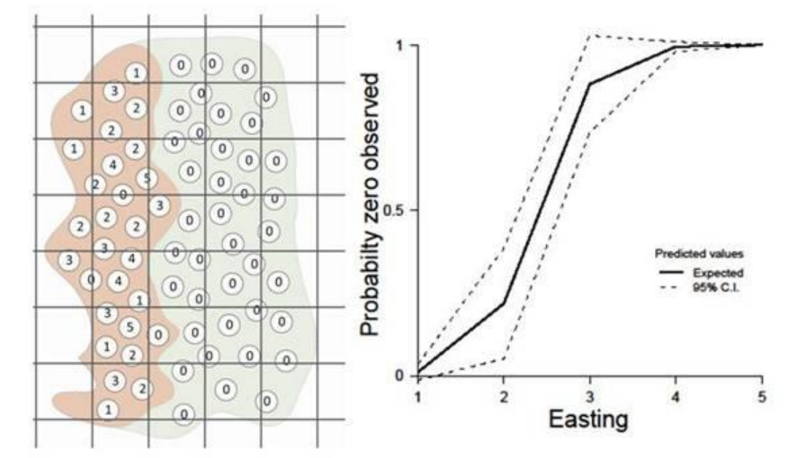
- Observer 2 tends to detect more infected hosts, where these are present.
- This detection heterogeneity can be modeled as a categorical term (Observer 1 vs. Observer 2). This is called a *detection covariate*.

Modeling the heterogeneous detection probability, using detection covariates, can be used to correct underreporting by *both* observers:



False negatives and underestimates have been reduced through state-space modeling

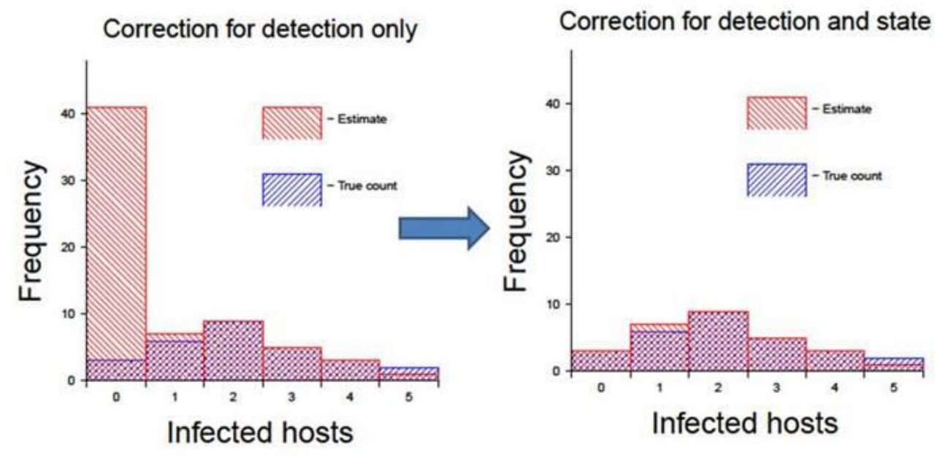
State-space modeling includes models for both detection and state, so lets add a model for heterogeneous abundance:



- We notice a potential relationship between spatial location, and the probability that no infection occurs.
- We can include the spatial term "Easting" (easterly change in longitude) in the model tier for the state (abundance of infected hosts).

12

The full model



- **Including heterogeneous detection and heterogeneous abundance allows us to make a precise estimate of our parameter-of-interest, the abundance of infected hosts!**

13

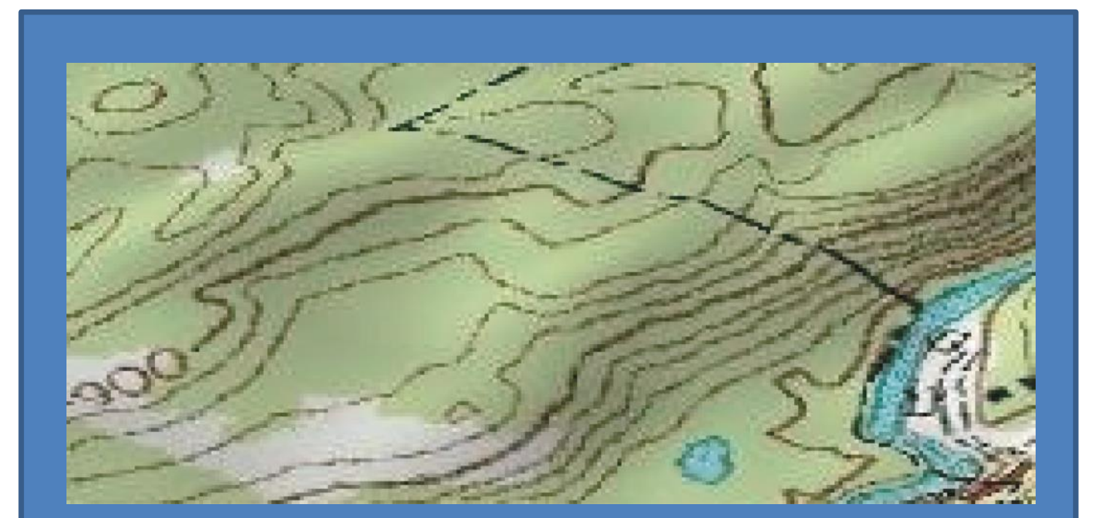
How It Would Work With Aerosol Sensor Data (simulation)

While state-space modeling has been applied to a wide variety of other surveillance data, we could find no instance in literature in which state-space models have been applied to bio-aerosol warning systems.

Methods:

We simulated data to emulate 30 Tactical Biological (TAC-BIO) Detectors or Biological Agent Warning Sensors **(**BAWS), using a binomial random number generator to simulate imperfect detection. Detectors were employed in pairs. Thirty detectors were used across 1000 simulations for each analytical method. TRM data included no within-pair detection heterogeneity, simulating conditions where samples are taken in rapid succession. MIO data included within-pair detection heterogeneity, simulating fusion of data from unequal observers or unequal sensor technology. Detection probability was set at 60% for TRM and was allowed to vary around 60% for MIO. Simple single-sample methods treated all detectors as independent. Single-sample methods, where undetected incidence is not estimated, are currently standard for most surveillance. We estimated the proportion of area affected using each method, then compared estimates to the "true" area used for the simulation. Data generation was performed using the mathematical programming language R 3.0.0(R Core Group 2013), and state-space models used the R library *unmarked* (Fiske and Chandler 2011).

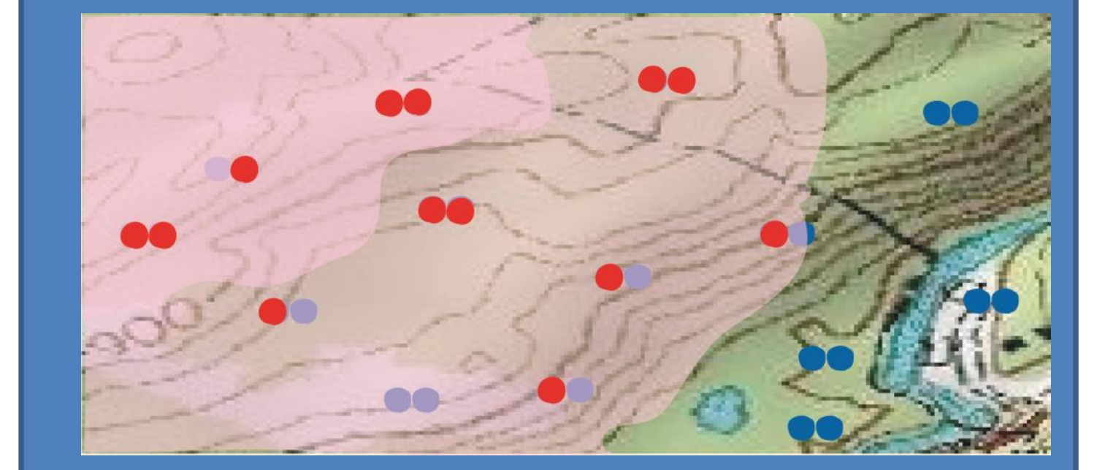
Setup:



Suppose we have some terrain to monitor

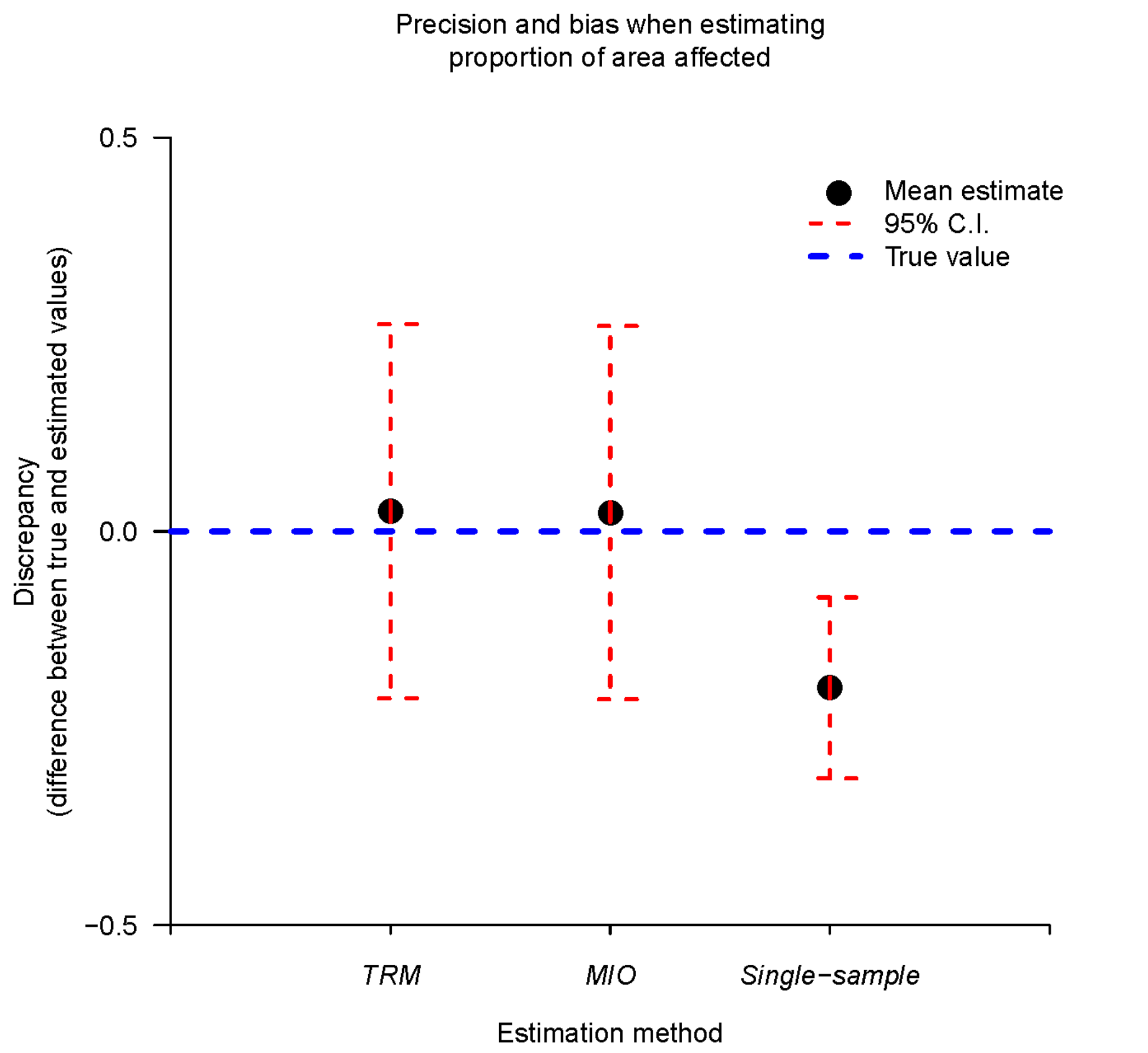
We deploy sensors in pairs along a grid or at random

An aerosol infiltrates the area



Triggering some, but not all the alarms

Results:



TRM and MIO models gave estimates that were nearly unbiased, that is, the mean of the estimates was very close to the true values. Confidence intervals were relatively broad, reflecting the need to collect larger samples for these models. Single-sample estimates were biased downwards in all cases due to underreporting.

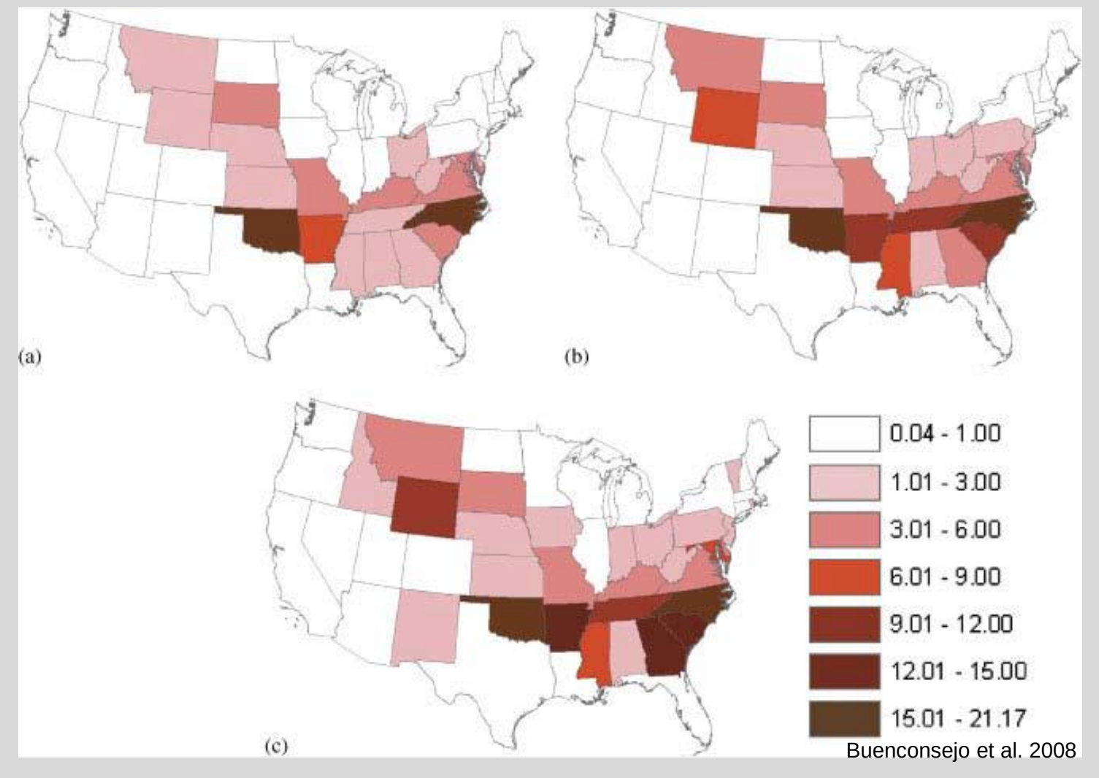
Conclusions:

State-space modeling is a promising method that can provide better estimates of affected areas than currently used single-sample methods. State space methods will require an increase in sampling frequency or intensity. However, adopting state-space methods for threat surveillance would reduce false-negative/positive alarms in sensor networks, allow data-fusion between sensors of unequal sensitivity, and provide a more robust and complete assessment of the threat environment.

In the Literature

State-space modeling originated as a method for estimating size and distribution of imperfectly observed populations (de Valpine and Hastings 2002), and has since been extended to cover a wide variety of estimation needs. State-space modeling has been particularly useful for estimating under-reporting in disease studies (Abubakar et al. 2010; McClintock et al. 2010), and attribution of false-positives in distribution estimates (Royle and Link 2006). Special modeling accomplishments that appear in literature and are especially relevant to risk assessment are shown below.

Correcting for under-reporting using fusion of two incomplete data sets



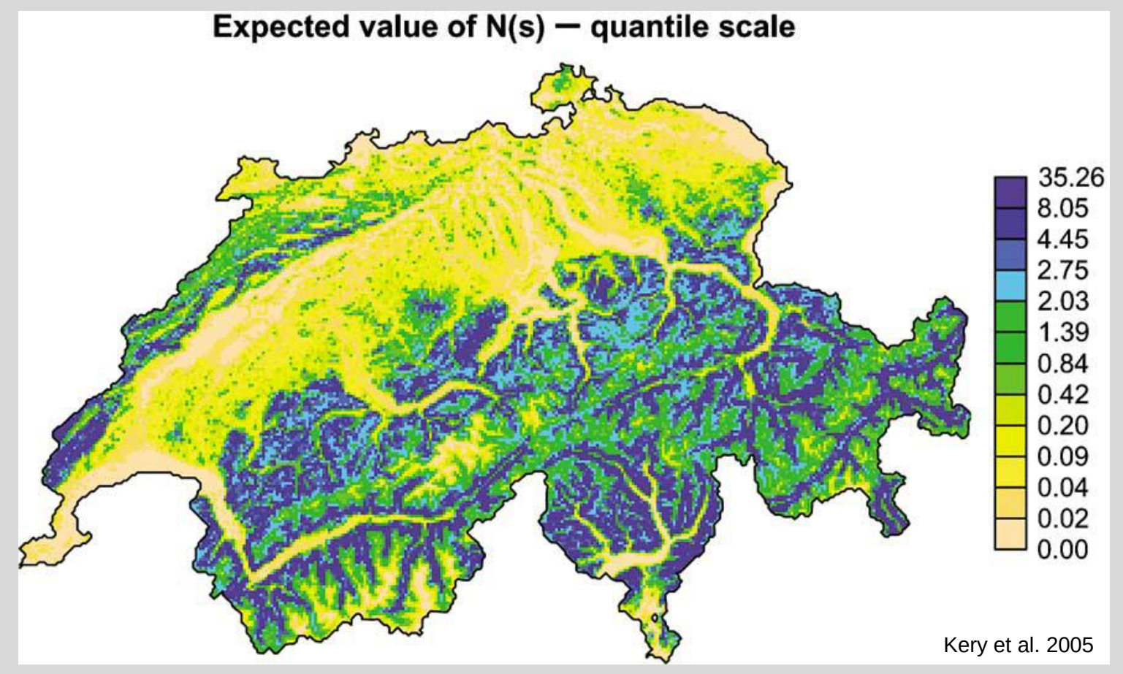
Buenconsejo et al. 2008

Buenconsejo et al. (2008) used a type of state-space model, called a spatial capture/recapture (SCR) model, to reduce estimation errors for density of Rocky Mountain spotted fever (RMSF) present in the Case Reporting Form (CRF:map a), and the National Electronic Telecommunications System for Surveillance (NETSS:map b).

Data fusion through SCR (map c) increases density estimates for RMSF over both single-source estimates.

SCR is not simply the union of CRF and NETSS, but includes an estimated count of cases missed in both archives!

Mapping of an imperfectly observed population using geo-referenced covariates



Kery et al. 2005

Kery et al. (2005) used a state-space abundance model to predict population sizes in imperfectly counted birds according to modeled associations with environmental co-variates archived in Geographic Information Systems (GIS).

State-space models allow co-variates such as forest cover that affect both state (abundance in this case) and detection to be un-confounded.

A map of the modeled population is then produced using GIS rasters.

Literature cited:

Abubakar, I., A. Bassili, A. Bierrenbach, E. Bloss, K. Floyd, P. Glaziou, R. Harris, K. Lonnroth, F. Mecatti, A. Pavli, C. Sismanidis, H. Timimi, M. Uplekar, R. van Hest, P. Glaziou, and K. Floyd. 2010. Assessing tuberculosis under-reporting through inventory studies. WHO Press, Geneva.

Buenconsejo, J., D. Fish, J. E. Childs, and T. R. Holford. 2008. A Bayesian hierarchical model for the estimation of two incomplete surveillance data sets. Statistics in Medicine **27**:3269–3285.

de Valpine, P., and A. Hastings. 2002. Fitting population models incorporating process noise and observation error, *Ecological Monographs* **72**, 57-76.

Fiske, I., and R. Chandler. 2011. unmarked: An {R} Package for Fitting Hierarchical Models of Wildlife Occurrence and Abundance. Journal of Statistical Software **43**: 1-23.

Kery, M, J.A. Royle, and H. Schmid. 2005. Modeling avian abundance from replicated counts using binomial mixture models. Ecological Applications. **15:** 1450-1461.

McClintock, B. T., J. Nichols, L. Bailey, D. MacKenzie, W. Kendall, and A. Franklin. 2010. Seeking a second opinion: uncertainty in disease ecology. Ecology Letters **13**: 659–674.

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Royle, J. A., and W. A. Link. 2006.Generalized site occupancy models allowing for false positive and false negative errors. Ecology, **87**: 835–841.

Acknowledgements: The authors thank the Defense Threat Reduction Agency/Joint Science and Technology Office for their assistance and funding of this work. The views expressed in this presentation are those of the authors and do not necessarily reflect the official policy or position of the Department of Defense or the U.S. Government.